

Four Approaches to Low-Resource Multilingual NMT: The Helsinki Submission to the AmericasNLP 2023 Shared Task

Ona de Gibert Raúl Vázquez Mikko Aulamo
Yves Scherrer Sami Virpioja Jörg Tiedemann

University of Helsinki, Dept. of Digital Humanities
{firstname.lastname}@helsinki.fi

Abstract

The Helsinki-NLP team participated in the AmericasNLP 2023 Shared Task with 6 submissions for all 11 language pairs arising from 4 different multilingual systems. We provide a detailed look at the work that went into collecting and preprocessing the data that led to our submissions. We explore various setups for multilingual Neural Machine Translation (NMT), namely knowledge distillation and transfer learning, multilingual NMT including a high-resource language (English), language-specific fine-tuning, and a system with a modular architecture. Our multilingual Model B ranks first in 4 out of the 11 language pairs.

1 Introduction

This paper presents the submission of the Helsinki-NLP team to the AmericasNLP 2023 Shared Task. The task consisted in developing Machine Translation (MT) systems for 11 indigenous languages of the Americas: Aymara (aym), Bribri (bzd), Asháninka (cni), Chatino (czn), Guarani (gn), Wixarika (hch), Nahuatl (nah), Hñähñu (oto), Quechua (Quy), Shipibo-Konibo (shp), and Rarámuri (tar). The AmericasNLP task has been running for two years: in 2021 (Mager et al., 2021) it was first introduced, and in 2022 it consisted of Speech-to-Text Translation (STT).¹ This year’s task is similar to the one held in 2021, but it includes an additional language (Chatino) and the use of the development set in training is not allowed. Our 2021 submission (Vázquez et al., 2021) reached the first rank in nine out of ten languages and serves as the baseline for this year’s task.

The 11 target languages involved in the task vary a lot in terms of “resourcedness”. On one side of the spectrum, there are languages like Quechua and Guarani with millions of native speakers, whereas on the other end, the variety of Hñähñu

¹<http://turing.iimas.unam.mx/americasnlp/st.html>

used in the development and test sets only has about 100 elder speakers.² Many of the target languages show dialectal variation, and some have different spelling norms and conventions. Furthermore, some datasets contain instances of code-switching with Spanish, and some of the languages are polysynthetic. All these factors make the task at hand particularly challenging.

A large part of our effort focuses on increasing the amount of parallel data for training. Building on our work for the 2021 shared task, we employ several strategies: mining, extraction and alignment of publicly available parallel resources, backtranslation of monolingual data (Sennrich et al., 2016), and data augmentation by pivoting through English (Xia et al., 2019).

On the modelling side, our winning 2021 submission was based on a multilingual (one-to-many) model that was pretrained mostly on the Spanish-to-English task and later fine-tuned on the low-resource indigenous languages. We keep this general approach in most of this year’s submissions, but provide some variations to this theme:

Model A uses knowledge distillation and transfer learning instead of training from scratch. In this context, we also experiment with different data labeling schemes.

Model B reproduces our 2021 setup with updated data.

Model C reimplements Model B’s strategy using OpusTrainer³ and introduces a language-specific fine-tuning step.

Model D uses a modular architecture in a multilingual setting with language-specific decoder modules.

²https://github.com/AmericasNLP/americasnlp2023/blob/main/data/information_datasets.pdf

³<https://github.com/hplt-project/OpusTrainer>

Our best-performing model is Model B. The collected data and our code are publicly available on our fork of the organizers’ Git repository.⁴

The rest of the paper is organised as follows. Section 2 provides a detailed description of our data collection and preparation efforts. Section 3 describes in detail the models presented. Section 4 outlines the results and, finally, section 5 concludes our work.

2 Data collection and preparation

Similar to our 2021 submission, we worked on finding relevant corpora from additional sources and cleaning and filtering them. We utilised the OpusFilter toolbox⁵ (Aulamo et al., 2020), which provides both ready-made and extensible methods for combining, cleaning, and filtering parallel and monolingual corpora. OpusFilter uses a configuration file that lists all the steps for processing the data; in order to make quick changes and extensions programmatically, we generated the configuration file with a Python script.

2.1 Data collection

We combined the data previously collected for our 2021 participation with some new resources. An overview of the resources, including references and URLs, is given in Table 4 in the appendix.

Organizer-provided resources The shared task organizers provided parallel datasets for training for all 11 languages. These datasets are referred to as *train* in this paper. For some of the languages (e.g., Ashaninka, Wixarika and Shipibo-Konibo), the organizers pointed participants to repositories containing additional data. We refer to these resources as *extra*. Furthermore, the organizers provided development (*dev*) and test (*test*) sets for all 11 language pairs of the shared task (Ebrahimi et al., 2023).

OPUS The OPUS corpus collection (Tiedemann, 2012) provides only few datasets for the relevant languages. We utilized the *GNOME*, *MozillaI10n* and *Ubuntu* corpora, which consist of localization files. Additionally, we made use of the *Tatoeba* and *Wikimedia* corpora, which have been recently updated on the OPUS website.⁶ These bitexts contain

⁴<https://github.com/Helsinki-NLP/amicasnlp2023-st>

⁵<https://github.com/Helsinki-NLP/OpusFilter>, version 2.6.

⁶<https://opus.nlpl.eu/>

384 sentence pairs for Aymara, 25233 for Guarani, 169 for Nahuatl and 1187 for Quechua parallel with Spanish.

To ensure collecting data only for the relevant languages, we ran language detection on the corpora. For language identification we used HeLI-OTS (Jauhiainen et al., 2022), which includes language models for Guarani, Nahuatl and Quechua. We kept only pairs where both the source and the target sentences are detected to be in the correct language. For the Spanish side, we also accepted sentences identified as other Romance languages, namely Catalan, Galician, French, Portuguese, Extremaduran and Occitan. For Aymara and Nahuatl, we chose to accept sentences where the detected language is not English or Spanish, as Aymara is not included in the language model and only a small proportion of sentences were detected to be Nahuatl. The language identification filtering leaves 320 sentence pairs for Aymara, 19751 for Guarani, 153 for Nahuatl and 718 for Quechua.

FLORES The FLORES-200 development and test sets (NLLB Team et al., 2022) cover Aymara, Guarani and Quechua. Since this is a multiparallel dataset, we paired the indigenous languages with their corresponding Spanish sentences. We concatenated the development and test sets and added them to our training data.

Bibles The JHU Bible corpus (McCarthy et al., 2020) covers all languages of the shared task with at least one Bible translation. When several Bibles were available for a given indigenous language, we scored them with a character 6-gram language model trained on the development sets and chose the Bible(s) with the lowest average cross-entropy scores. We paired them with the available Spanish Bibles using the product method in OpusFilter to randomly take at most 3 different versions of the same sentence (skipping empty and duplicate lines).⁷

Legal texts, educational material and news In 2021, we collected constitutions and laws of various Latin American countries with their translations into indigenous languages. We expanded this collection by adding the Chatino–Spanish Mexican constitution. We also added the Universal Declaration of Human Rights (UDHR) where avail-

⁷We sampled three Spanish sentences when there was a single Bible version for the indigenous language, two for 2–3 versions, and one for more than three versions.

able in the Universal Declaration of Human Rights Translation Project.⁸ Furthermore, we extracted Nahuatl and Bribri educational material as well as Guaraní parallel news items from PDF documents and websites. The document and sentence alignment was done semi-automatically using source-specific heuristics and the hunalign⁹ (Varga et al., 2005) tool. We provide a script in our repository to replicate these data gathering and alignment procedures.¹⁰

Spanish–English data All submitted models take advantage of abundant parallel data for Spanish–English. The resources come from OPUS (Tiedemann, 2012) and include the following sources: *OpenSubtitles*, *Europarl*, *GlobalVoices*, *News-Commentary*, *TED2020*, *Tatoeba*, *bible-uedin*. The Spanish–English *WMT-News* corpus, also from OPUS, is used for validation.

2.2 Back-translations of monolingual data

The organizers also provided some monolingual resources for some indigenous languages. We also obtained monolingual Wikipedia dumps for some languages through the Tatoeba Translation Challenge project (Tiedemann, 2020). We used the 2021 reverse Model B to translate these resources to Spanish (thereby fixing the processing for Quechua reported in the 2021 paper).


2.3 Pivot translations of English-aligned data

Some parallel datasets provided by the organizers or available on OPUS were aligned with English. Furthermore, the No Language Left Behind (Costajussà et al., 2022) project released training data for Aymara–English and Guaraní–English. We used a publicly available English-to-Spanish MT system from the OPUS-MT project¹¹ to translate the English side to Spanish in order to constitute additional Spanish–Indigenous data.

2.4 Data normalization, cleaning and filtering

We noticed that some of the corpora in the same language used different orthographic conventions

and had other issues that would hinder NMT model training. We applied various data normalization and cleaning steps to improve the quality of the data, with the goal of making the training data more similar to the development data (which we expected to be similar to the test data).

For Bribri, Raramuri and Wixarika, we found normalization scripts or guidelines on the organizers’ Github page or sources referenced therein (cf.  entries in Table 4). We reimplemented them as custom OpusFilter preprocessors. For Chatino, we implemented a preprocessor that normalized the tone characters variations in the different datasets.

The organizer-provided training sets for Bribri, Hñähñu, Nahuatl, and Raramuri were originally tokenized. We detokenized these corpora with the Moses detokenizer supported by OpusFilter, using the English patterns. Finally, for all datasets, we applied OpusFilter’s `WhitespaceNormalizer` preprocessor, which replaces all sequences of whitespace characters with a single space.

We filtered some of the datasets using predefined filters from OpusFilter. Not all filters were applied to all languages; instead, we selected the appropriate filters based on manual observation of the data and the proportion of sentences removed by the filter. Appendix A describes the filters in detail.

2.5 Data tagging

Since all our models are multilingual models with several target languages, we include a **target language tag** at the beginning of the source sentence. Furthermore, we add two more tags: variant tags and quality tags.

Variant tags represent the different variants of a particular language and they were inferred either from the documentation of the data source or from a manual inspection focusing on the character set of the specific text. In the end, we only used variant tags for two languages: Chatino and Quechua. The `<default>` variant is always the variant of the development and test sets. Besides the `<default>` variant, for Chatino we define the `<plain>` variant, which does not use tones. It is important to mention that 95% of our training data for Chatino belongs to the `<plain>` variant. For Quechua, the development and test data is in Ayacucho Quechua (quy), whereas other data are in Cuzco Quechua or a Bolivian variety of Quechua. We define the variant labels `<quz>` and `<quh>` for the latter two.

Quality tags refer to the origin of the data:

⁸<https://www.ohchr.org/en/human-rights/universal-declaration/universal-declaration-human-rights/about-universal-declaration-human-rights-translation-project>

⁹<https://github.com/danielvarga/hunalign>

¹⁰under `data/getdata2023.py`

¹¹We used the `opusTCv20210807+bt_transformer-big_2022-03-13` model from <https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-spa>.

<default> for relatively clean data sources, <noisy> for unreliable data sources or with noisy sentence alignment, <bt> for back-translations, and <bible> for Bibles. The statistics of the quality tags for the training corpora are provided in subsection 2.8.

If not specified otherwise, all tags are used during the training phase. When generating test translations, we use the language tag, followed by the default variant and quality tags.

2.6 Concatenation and deduplication

After tagging, the different training sets were concatenated, and all exact duplicates were removed from the data using OpusFilter’s duplicate removal step. Note that because of the language variant tags, some duplicates marked as different variants may have remained.

For the Spanish–English data, duplicates were removed separately from the OpenSubtitles part and the rest of the data.

2.7 Data postprocessing

We apply data postprocessing steps for two target languages: Chatino and Hñähñu.

Chatino has a tonal structure, where each word is tagged at the end with a superscript tone character (^AB^cE^fG^HI^JK), for example: *Kyqya^A no^A shtya^H renq^J 2/2022-CC qo^E 4/2022-CC*. Sometimes, the character ^J can also be found within a word. A manual inspection of the results allowed us to see that our models were not producing the superscript characters, presumably due to Unicode normalization performed during subword segmentation with SentencePiece. Therefore, we opted for substituting the characters in the character set mentioned above by their superscript counterparts if they were found at the end of a token. For ^J, we replaced all occurrences regardless of their position.

Regarding **Hñähñu**, organizers already acknowledge that the training variant (Valle del Mezquital) is a different one from the development and test sets (Ñûhmû de Itxenco), a severely endangered variant spoken by less than 100 people. The training data did not contain any sample from the development and test set variant, having some characters in the training data that never appear in the development set. In consequence, we chose to substitute all occurrences of the character set that only appear in the training data, by their non-diacritic counterpart. For example, *ë* becomes *e*, *è* becomes *e* and *ě* becomes *e*. The full character substitution can be

consulted in our GitHub repository.

2.8 Data sizes

Table 1 shows the sizes of the used datasets. *train* refers to the official training data and *extra* to all other datasets except the Bibles. The data sizes are listed separately before and after filtering, as well as after concatenation and duplicate removal (*combined*). There is a difference of almost two orders of magnitude between the smallest (czn) and largest (quy) combined training data sets. Including the Bibles data (*bibles*) evens out the situation a bit, but Quechua has still significantly more data than any of the other languages. The development sets comprise 500–1000 sentences for each of the languages.

As discussed in subsection 2.5, we use different quality tags for different data sources. Table 1 also shows the amount of the different tags in the *combined* set. In addition, <bible> was used always for *bibles*.

Finally, Table 2 shows the sizes of the Spanish–English datasets before and after filtering. Model A uses different data than models B, C and D; see section 3 for details.

3 Models

We tested four major model configurations, which we refer to as A, B, C and D. All models are multilingual neural MT (NMT) models and include the Spanish–English translation task in some form. Models B and C also include language-specific fine-tuning steps. All models are based on the Transformer architecture (Vaswani et al., 2017). Models A and C are trained using the MarianNMT Toolkit (Junczys-Dowmunt et al., 2018), while B and D are implemented with OpenNMT-py 2.0 (Klein et al., 2020). All models were trained on a single GPU, except Model D, which was trained on 4 GPUs.

We use subword SentencePiece segmentation (Kudo and Richardson, 2018) for the training data. We train a shared vocabulary for all languages with size 32k that is used in all the models. Further details of the configurations are listed in Appendix B.

3.1 Model A

Model A is a multilingual one-to-many model based on knowledge distillation (Kim and Rush, 2016), where you distill a smaller student model from a powerful teacher; and transfer learning (Zoph et al., 2016), where you train a parent model

	Data type	train		extra		combined (train+extra)				bibles
		none	filtered	none	filtered	filtered+deduplicated				filtered
	Quality tag					all	<default>	<noisy>	<bt>	<bible>
Ashaninka	cni	3,883	3,878	13,195	8,593	12,448	3,855	–	8,593	23,321
Aymara	aym	6,531	6,039	34,551	27,265	33,136	22,380	288	10,468	92,082
Bribri	bzd	7,508	7,490	659	588	7,853	7,519	334	–	23,103
Chatino	czn	357	354	4,841	4,798	4,804	4,804	–	–	47,570
Guarani	gn	26,032	26,012	82,703	72,597	86,698	36,435	16,833	33,430	23,687
Hñahñu	oto	4,889	4,888	9,013	8,593	13,401	13,331	70	–	23,849
Nahuatl	nah	16,145	15,863	26,892	22,558	35,360	27,839	1,473	6,048	47,674
Quechua	quy	125,008	109,372	261,055	209,814	306,999	268,020	617	38,362	123,829
Raramuri	tar	14,720	14,495	2,255	2,194	16,529	16,529	–	–	23,678
Shipibo-Konibo	shp	14,592	14,553	40,317	36,029	49,428	29,977	78	19,373	47,638
Wixarika	hch	8,966	8,960	3,165	2,932	11,784	11,518	–	266	23,867

Table 1: Numbers of segment pairs used for training (*train*: official training set provided by the organizers; *extra*: additional training data collected by the organizers and us, including back-translations and pivoted data but excluding Bibles; *bibles*: generated Bible data segments). The table also shows the effect of filtering and deduplication, as well as the repartition of data over the different quality tags (<default> for relatively clean data sources, <noisy> for unreliable data sources or with noisy sentence alignment, and <bt> for back-translations).

	news		opensubs		bibles	dev
	none	filtered+deduped	none	filtered+deduped	filtered	none
	<default>		<noisy>		<bible>	<default>
Model A	–	–	61,434,251	26,158,993	–	9,122
Models B, C, D	3,761,249	3,346,060	61,447,674	20,343,327	61,198	14,522

Table 2: Spanish–English dataset sizes: *news* is the combination of other training corpora (Europarl, GlobalVoices, News-Commentary, TED2020, Tatoeba) than OpenSubtitles and Bibles. The *dev* set for Model A consists of Spanish side of the official development sets machine-translated to English, and the WMT-News corpus for the other models.

on a high-resource pair and then continue training a child model on the low-resource data.

Regarding transfer learning, we train a parent model on a high-resource language pair (*es-en*) and then we continue training on the indigenous languages’ data. Furthermore, for the *es-en* parent model, we apply knowledge distillation. We distill a *es-en* system from the No Language Left Behind (NLLB) model¹² (Costa-jussà et al., 2022) by simply training a new model on NLLB translated data from Spanish into English. The rationale behind this decision is to benefit from the advantages of a large pretrained NMT model while optimizing its size to enable effective fine-tuning.

In contrast to the other models, we exclusively use the OpenSubtitles dataset for Spanish–English training. This dataset consists of relatively brief sentences discussing general subjects. The motivation to use only this dataset was based on an examination of the development sets, which exhibited similar content characteristics. For development, we translate the source Spanish counterpart of the development sets provided by the organizers into English with the NLLB model with the hope that the distilled model will overfit to its teacher’s distributions.

For the child model, we experiment with different data labeling schemes and submit three different versions:

- A.1: Parent model fine-tuned on indigenous data with all tags.
- A.2: Parent model fine-tuned on indigenous data without quality tags (keeping only the language and variant tags)
- A.3: Ensemble model of A.1 and A.2

3.2 Model B

Model B is a multilingual one-to-many model that reproduces the Model B setup from 2021 with updated training data.

The training takes place in three phases. In the first phase, the model is trained on 91% of Spanish–English data and 9% of data coming from the indigenous languages. The two English sets, *news* and *opensubs*, were assigned the same weight to avoid overfitting on subtitle data. In the second phase, the proportion of Spanish–English data is

¹²We use the NLLB-200’s 3.3B variant as the teacher. <https://huggingface.co/facebook/nllb-200-3.3B>

reduced to 37%, with the remainder sampled to equal amounts from the indigenous languages.

We train the first phase for 100k steps and pick the best intermediate savepoint according to the English validation set, which occurred after 80k steps. We initialize phase 2 with this savepoint and continue training until 200k steps. We then pick the five most promising savepoints based on the accuracy of the concatenated development sets, and select the best out of these five for each target language separately.

Starting from these savepoints, we added a third phase with language-specific finetuning, using 40% of English data and 60% of the individual target-language data. We trained these models for an additional 12k steps and selected the best intermediate savepoint. However, language-specific finetuning only increased the results for Ashaninka, Guarani and Raramuri. For the other languages, we used the best model savepoint from the second phase.

3.3 Model C

Model C is a set of 11 different language-specific models following the same strategy as Model B, trained with OpusTrainer.¹³ OpusTrainer is a tool for curriculum learning, especially designed for multilingual scenarios, since it allows to specify the desired mixture of datasets from different language sources.

Similarly to Model B, the training takes place in three phases. We train our models with all the available data for all language pairs with the following configuration: (1) First, we train for one epoch with 90% of the *es-en* data and 10% of indigenous data, coming from each of the 11 indigenous languages. (2) Then, we train two epochs with a 50/50 distribution. Finally, (3) we add a language-specific fine-tuning step, where we train with a distribution of 10% of *es-en* data, 10% of *es-indigenous* and 80% of the desired language until convergence with early-stopping.

For inference, we ensemble the last four checkpoints with different combinations (1, 1-2, 1-2-3, 1-2-3-4) for each model. We select the best ensemble approach for each language pair based on the development set scores.

3.4 Model D

Model D is a multilingual modular sequence-to-sequence Transformer model (Vázquez et al., 2020;

¹³<https://github.com/hplt-project/OpusTrainer>

Escolano et al., 2021). It is trained to perform Spanish-to-many translation, as well as a denoising auto-encoding objective (Lewis et al., 2020) for each of the 11 indigenous languages as well as English. Each model consists of 12 layers: a 6-layer Spanish encoder and decoders that share s layers followed by $6 - s$ language-specific layers. We trained distinct models with $s = 1, 2, 3$. Model D is set to $s = 1$ since it outperformed the others with respect to ChrF scores in the development set. Training details are given in Appendix B.

4 Results

Our results are shown in Table 3 with the official automatic evaluation metric, ChrF (Popović, 2015). We also include the results of this year’s baseline and the best of the contenders for each of the target languages.

The baseline turned out to be quite hard to beat: for five languages (*hch, nah, oto, shp, tar*), the best submission was less than 2 ChrF points above the baseline. The competition among participants was also very tight this year: for the same five languages, there is less than 1 ChrF point difference between the first and second participant. Differences of less than 2 ChrF points can be observed for two additional languages (*cni, gn*). We believe that conducting significance testing to compare the participants’ results would be beneficial in this scenario.

Regarding our models, Model B is our clear best-performing system. It reached first rank on 4 out of the 11 language pairs and third rank on two other occasions. Model B consistently outperformed all our other models. Its good performance can be attributed to its pre-training phase on Spanish-English data including a small percentage of the indigenous data. For this model, we also focused our efforts in checkpoints’ selection. Further analysis will be required to investigate the performance differences between our models B and C, which used the same overall setup but show various minor differences in terms of toolkits, hyperparameters and curriculum definition.

The variants of Model A perform very similarly to each other, although removing the quality tags (A.2) leads to a significant increase for *es-shp*. Comparing models A and model B, our results indicate that training a multilingual model jointly from scratch is more beneficial than transfer learning approaches.

Model C seems to be on par with models A, although it works particularly well for *es-czn*. With Model C, we expected that language-specific fine-tuning would boost results. If we compare models B and C, our results match previous research, where it is stated that low-resource translation benefits from jointly-trained multilingual models (Johnson et al., 2017).

Finally, while Model D works well for *es-shp*, outperforming models C and A.2, we observe that in general it yields poor results. Nonetheless, we decided to use it anyway to test it in a real use case. Specifically for Model D, we were interested in testing the knowledge transfer capabilities of modular systems in low-resource multilingual scenarios. Indeed, these systems have demonstrated efficient transfer learning properties (Escolano Peinado, 2022). However, in this set of experiments, Model D lags behind our other non-modular systems for all other languages, indicating that perhaps the data available to train the language-specific modules was insufficient or that the parameter sharing strategies we chose were not optimal. In our experiments we also noticed that the modular systems ignore the variant and quality tags, which hampers their performance due to the imbalance of training resources. This can be seen in the case of *es-czn*, where the model is unable to learn the variant of the test set due to the unbalanced amount of that variant in the training data (only 5%).

5 Conclusions

In this paper, we have presented our contribution to the AmericasNLP 2023 Shared Task. We have described our efforts in terms of data collection and processing. We presented our 6 submissions to the task for all language pairs. We explore various setups for multilingual NMT, including knowledge distillation, transfer learning, multilingual NMT with English, language-specific fine-tuning, and a multilingual modular system.

Our strongest system follows the same architecture as our winning submission in 2021, which was used as the baseline for this year. There are two main differences between our current submission and the baseline:

- Additional training data: the amount of added resources varies across the languages, and not all of our collection efforts seem to have paid off. While results improved substantially for Guarani, no significant improvements could

Data	Model	Run	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar	Average
dev	baseline		32.7	23.8	26.8	–	31.1	29.9	29.8	14.7	33.8	31.7	19.6	27.39
	A.1	1	36.0	19.6	26.0	13.5	34.8	29.3	27.6	13.1	35.9	22.4	18.4	25.15
	A.2	2	35.3	18.2	26.9	13.0	34.8	28.8	27.8	13.1	35.9	27.2	18.1	25.37
	A.3	4	36.4	19.7	26.0	13.5	36.0	29.3	29.0	13.2	36.4	23.7	18.0	25.56
	B	6	37.2	21.9	29.2	17.0	38.3	31.7	31.2	14.5	34.0	34.3	20.3	28.15
	C	3	34.8	18.9	26.5	14.4	35.1	29.0	27.3	13.2	33.9	21.5	18.6	24.84
	D	5	23.1	10.4	20.5	7.0	29.7	19.8	21.4	9.4	26.5	22.5	13.3	18.51
test	baseline		28.30	16.50	25.80	–	33.60	30.40	26.60	14.70	34.30	32.90	18.40	26.15
	best contender		36.24	26.08	29.98	39.97	39.34	32.25	27.33	14.81	39.52	33.43	18.74	–
	A.1	1	32.31	20.18	25.18	21.89	37.23	29.47	23.96	13.93	36.22	19.66	17.67	25.25
	A.2	2	31.98	19.19	25.99	21.67	36.60	29.48	25.61	14.23	36.49	25.41	17.45	25.83
	A.3	4	32.52	20.28	25.14	22.61	37.97	29.90	25.82	14.11	37.19	20.51	17.04	25.74
	B	6	33.44	22.45	28.41	32.07	40.42	32.34	26.87	15.30	33.29	33.35	19.15	28.83
	C	3	32.34	20.06	25.62	26.73	37.38	30.76	23.72	13.92	34.97	19.68	18.43	25.78
	D	5	21.86	11.16	19.60	7.17	31.15	21.01	19.87	10.66	27.72	22.85	12.92	18.72

Table 3: ChrF scores for the six submissions, computed on the development and test set. The Run column provides the numeric IDs with which our submissions are listed in the overview paper. In addition, we provide the baseline and the best competitor scores for each target language.

be observed for Nahuatl and Quechua. For Bribri, the model generalizes better to the test set than in 2021, but is still far behind the best contender.

- Inclusion of variant and quality tags: the experiments with Model A suggest that variant and quality tags can help, but that our current attribution of tags was not optimal. It could be promising to base the tags on more objective criteria like character and word overlap or alignment quality.

These two additions have allowed us to beat our own baseline.

Acknowledgements

This work was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 771113.

This work was also supported by the HPLT project which has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070350.

References

Željko Agić and Ivan Vulić. 2019. [JW300: A wide-coverage parallel corpus for low-resource languages](#).

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. [OpusFilter: A configurable parallel corpus filtering toolbox](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

David Brambila. 1976. *Diccionario Raramuri – Castellano (Tarahumara)*. Obra Nacional de la Buena Prensa, México.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. [No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. [Development of a Guarani - Spanish parallel corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Christos Christodoulopoulos and Mark Steedman. 2015. [A massively parallel corpus: the Bible in 100 languages](#). *Language Resources and Evaluation*, 49:375–395.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling

- human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. *Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar*. <http://www.lengamer.org/publicaciones/diccionarios/>.
- Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Enora Rice, Cynthia Montañó, John Ortega, Shruti Rijhwani, Alexis Palmer, Rolando Coto-Solano, Hilari Cruz, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics.
- Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. *Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.
- Carlos Escolano Peinado. 2022. *Learning multilingual and multimodal representations with language-specific encoders and decoders for machine translation*. *Ph.D. Thesis*, UPC, Departament de Teoria del Senyal i Comunicacions.
- Isaac Feldman and Rolando Coto-Solano. 2020. *Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. *Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo*. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.
- Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. *Axolotl: a web accessible parallel corpus for Spanish-Nahuatl*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).
- Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. *HeLI-OTS, off-the-shelf language identifier for text*. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 3912–3922, Marseille, France. European Language Resources Association.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. *Google’s multilingual neural machine translation system: Enabling zero-shot translation*. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. *Marian: Fast neural machine translation in C++*. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Yoon Kim and Alexander M. Rush. 2016. *Sequence-level knowledge distillation*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. *The OpenNMT neural machine translation toolkit: 2020 edition*. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.
- Taku Kudo and John Richardson. 2018. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. *Probabilistic finite-state morphological segmenter for wixarika (huichol) language*. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.
- Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. *Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas*. In

- Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.
- Jesús Manuel Mager Hois, Carlos Barron Romero, and Ivan Vladimir Meza Ruíz. 2016. Traductor estadístico wixarika - español usando descomposición morfológica. *COMTEL*, 6.
- Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. [The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.
- Elena Mihas. 2011. *Añaani katonkosatzi parenini, El idioma del alto Perené*. Milwaukee, WI: Clarks Graphics.
- Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. [A continuous improvement framework of machine translation for Shipibo-konibo](#). In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).
- John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. [Overcoming resistance: The normalization of an Amazonian tribal language](#). In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. [Parallel Global Voices: a collection of multilingual corpora with citizen media stories](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Noam Shazeer and Mitchell Stern. 2018. [Adafactor: Adaptive learning rates with sublinear memory cost](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2005)*, pages 590–596.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, 46(2):387–424.
- Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. [The Helsinki submission to the AmericasNLP shared task](#). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.
- Raúl Vázquez, Umut Sulubacak, and Jörg Tiedemann. 2019. [The University of Helsinki submission to the WMT19 parallel corpus filtering task](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 294–300, Florence, Italy. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. [Generalized data augmentation for low-resource translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. [Transfer learning for low-resource neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A OpusFilter settings

The following filters were used for the training data except for back-translated data, Bibles and the OpenSubtitles data for Model A:

- LengthFilter: Remove sentences longer than 1000 characters. Applied to Aymara, Chatino, Nahuatl, Quechua, Raramuri.
- LengthRatioFilter: Remove sentences with character length ratio of 4 or more. Applied to Ashaninka, Aymara, Chatino, Guarani, Hñähñu, Nahuatl, Quechua, Raramuri, Wixarika.
- CharacterScoreFilter: Remove sentences for which less than 90% characters are from the Latin alphabet. Applied to Aymara, Quechua, Raramuri.
- TerminalPunctuationFilter: Remove sentences with dissimilar punctuation; threshold -2 (Vázquez et al., 2019). Applied to Aymara, Quechua.
- NonZeroNumeralsFilter: Remove sentences with dissimilar numerals; threshold 0.5 (Vázquez et al., 2019). Applied to Aymara, Quechua, Raramuri, Wixarika.

The Bribri and Shipibo-Konibo corpora seemed clean enough that we did not apply any filters for them.

After generating the Bible data, we noticed that some of the lines contained only a single 'BLANK' string. The segments with these lines were removed afterwards.

From the provided monolingual datasets, we filtered out sentences with more than 500 words.

The back-translated data was filtered with the following filters:

- LengthRatioFilter with threshold 2 and word units
- CharacterScoreFilter with Latin script and threshold 0.9 on the Spanish side and 0.7 on the other side
- LanguageIDFilter with a threshold of 0.8 for the Spanish side only.

The OpenSubtitles data for Model A was filtered with the following filters:

- LengthRatioFilter with threshold of 3 and word units.
- CharacterScoreFilter with Latin script and threshold 0.75 on both sides.
- AlphabetRatioFilter with a default threshold of 0.75.
- LongWordFilter with a default maximum length of 40.
- AverageWordLengthFilter with default values of minimum length of 2 and maximum length of 20.

B Hyperparameters

Models A use a 6-layered Transformer with 8 heads, 512 dimensions in the embeddings and 2,048 dimensions in the feed-forward layers. The batch size is 1,000 sentence-pairs. The Adam optimizer is used with $\beta_1=0.9$ and $\beta_2=0.98$. The models are trained until convergence with early-stopping on development data after ChrF has stalled 10 times.

Model B uses a 8-layered Transformer with 16 heads, 1,024 dimensions in the embeddings and 4,096 dimensions in the feed-forward layers. The batch size is 9,200 tokens in phase 1 and 4,600 tokens in phase 2, with an accumulation count of 4. The Adam optimizer is used with $\beta_1=0.9$ and $\beta_2=0.997$. The Noam decay method is used with a learning rate of 2.0 and 16000 warm-up steps. Subword sampling is applied during training (20 samples, $\alpha = 0.1$). As a post-processing step, we removed the <unk> tokens from the outputs of Model B.

Model C uses a 6-layered Transformer with 8 heads, 512 dimensions in the embeddings and 2,048 dimensions in the feed-forward layers. The batch size is 1,000 sentence-pairs. The Adam optimizer is used with $\beta_1=0.9$ and $\beta_2=0.98$.

Model D was trained for a total of 150K steps to minimize the negative log-likelihood of the target translation. We accumulate gradients over all translation directions before back-propagation, using AdaFactor (Shazeer and Stern, 2018) with learning rate of 3.0. We trained the model on 4 AMD MI100 GPUs for ~ 48 hrs. The 8-headed Transformer layers have 512 dimensions in the self attention and 2,048 in the feed forward sub-layers.

Aymara aym	✿	GlobalVoices (Tiedemann, 2012; Prokopidis et al., 2016)
	☆	BOconst: https://www.kas.de/c/document_library/get_file?uuid=8b51d469-63d2-f001-ef6f-9b561eb65ed4&groupId=288373
	★	FLORES-200: https://github.com/facebookresearch/flores
	★♻️	NLLB-MD: https://github.com/facebookresearch/flores
	★	OPUS: Mozilla-I10n, wikimedia (Tiedemann, 2012)
	★	UDHR: https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection
	★♻️	GlobalVoices (en-aym) (Tiedemann, 2012; Prokopidis et al., 2016)
	☆🔄	OPUS: Wikipedia (Tiedemann, 2020)
	📖	<i>ayr-x-bible-2011-v1</i>
Bribri bzd	✿	(Feldman and Coto-Solano, 2020)
	★	MEP: https://mep.go.cr/educatico/minienciclopedias-pueblos-indigenas
	★	IUCN: https://portals.iucn.org/library/sites/library/files/documents/2016-071.pdf
	📖	<i>bzd-x-bible-bzd-v1</i>
	⚙️	https://github.com/AmericasNLP/americasnlp2021/blob/main/data/bribri-spanish/orthographic-conversion.csv
Ashaninka cni	✿	https://github.com/hinantin/AshaninkaMT (Ortega et al., 2020; Cushimariano Romano and Sebastián Q., 2008; Mihas, 2011)
	☆🔄	ShaShiYaYi (Bustamante et al., 2020): https://github.com/iapucp/multilingual-data-peru
	📖	<i>cni-x-bible-cni-v1</i>
Chatino czn	✿	https://scholarworks.iu.edu/dspace/handle/2022/21028
	★	MXconst: https://constitucionenlenguas.inali.gob.mx/
	★♻️	CTP-ENG: https://github.com/AmericasNLP/americasnlp2023
	📖	<i>cta-x-bible-cta-v1, ctp-x-bible-ctp-v1, cya-x-bible-cya-v1</i>
Guarani gn	✿	(Chiruzzo et al., 2020)
	★	PYconst: http://ej.org.py/principal/constitucion-nacional-en-guarani/
	★	News: https://spl.gov.py/es/index.php/noticias & https://www.spl.gov.py/gn/index.php/marandukuera
	★	Jojajovai: https://github.com/pln-fing-udelar/jojajovai
	★	FLORES-200: https://github.com/facebookresearch/flores
	★♻️	NLLB-seed: https://github.com/facebookresearch/flores
	★	UDHR: https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection

(Continues on next page)

Guarani (cont.)	★	OPUS: GNOME, Mozilla-I10n, Tatoeba, Ubuntu, wikimedia (Tiedemann, 2012)
	☆ ↻	OPUS: Wikipedia (Tiedemann, 2020)
	📖	<i>gug-x-bible-gug-v1</i>
Wixarika hch	🌟	https://github.com/pywirrarika/wixarikacorpora (Mager et al., 2018)
	☆	MXconst: https://constitucionenlenguas.inali.gob.mx/
	☆	corpora.wixes, paral_own, segcorpus.wixes: https://github.com/pywirrarika/wixarikacorpora
	☆ ↻	social.wix: https://github.com/pywirrarika/wixarikacorpora
	📖	<i>hch-x-bible-hch-v1</i>
	⚙️	https://github.com/pywirrarika/wixnlp/blob/master/normwix.py (Mager Hois et al., 2016)
Nahuatl nah	🌟	Axolotl (Gutierrez-Vasques et al., 2016)
	☆	MXConst: https://constitucionenlenguas.inali.gob.mx/
	★	Educational: https://nawatl.com/category/textos/
	★	Dict: https://nahuatl.wired-humanities.org/
	★	Short stories: https://nahuatl.org.mx/cuentos-nahuatl-14-ejemplares-para-descargar/
	★	INPI monograph: https://www.gob.mx/inpi/documentos/monografia-nacional-los-pueblos-indigenas-de-mexico & https://www.gob.mx/inpi/documentos/libros-en-lenguas-indigenas
	★	UDHR: https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection
	★	OPUS: Tatoeba, wikimedia (Tiedemann, 2012)
	☆ ↻	OPUS: Wikipedia (Tiedemann, 2020)
	📖	<i>azz-x-bible-azz-v1, ncj-x-bible-ncj-v1, nhi-x-bible-nhi-v1</i>
	Hnähñu oto	🌟
☆		MXConst: https://constitucionenlenguas.inali.gob.mx/
★		Dictionary: http://xixona.dlsi.ua.es/~fran/ote-spa.tsv
★		UDHR: https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection
📖		<i>ote-x-bible-ote-v1</i>
Quechua quy	🌟	JW300 (quy+quz) (Agić and Vulić, 2019)
	☆	MINEDU, dict_misc: https://github.com/AmericasNLP/americasnlp2021/tree/main/data/quechua-spanish
	☆	PEconst: https://www.wipo.int/edocs/lexdocs/laws/qu/pe/pe035qu.pdf

(Continues on next page)

Quechua (<i>cont.</i>)	☆	BOconst: https://www.kas.de/documents/252038/253252/7_dokument_dok_pdf_33453_4.pdf/9e3dfb1f-0e05-523f-5352-d2f9a44a21de?version=1.0&t=1539656169513
	★	UDHR (3 versions): https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection
	★	FLORES-200: https://github.com/facebookresearch/flores
	★ 🔄	JW300 (en–quy, en–quz) (Agić and Vulić, 2019)
	★	OPUS: GNOME, Mozilla-I10n, Tatoeba, Ubuntu, wikimedia (Tiedemann, 2012)
	☆ ↻	OPUS: Wikipedia (Tiedemann, 2020)
	📖	<i>quy-x-bible-quy-v1, quz-x-bible-quz-v1</i>
Shipibo-Konibo shp	🌟	(Galarreta et al., 2017; Montoya et al., 2019)
	☆	Educational, Religious: http://chana.inf.pucp.edu.pe/resources/parallel-corpus/
	★	LeyArtesano: https://cdn.www.gob.pe/uploads/document/file/579690/Ley_Artesano_Shipibo_Konibo_baja__1_.pdf
	★	Tsanas: http://chana.inf.pucp.edu.pe
	★	Covid19: https://github.com/iapucp/covid19-multilingue-peru
	★	UDHR: https://searchlibrary.ohchr.org/search?ln=en&cc=UDHR+Translation+Collection
	☆ ↻	ShaShiYaYi (Bustamante et al., 2020): https://github.com/iapucp/multilingual-data-peru
📖	<i>shp-SHPTBL</i>	
Raramuri tar	🌟	(Brambila, 1976)
	☆	MXConst: https://constitucionenlenguas.inali.gob.mx/
	📖	<i>tac-x-bible-tac-v1</i>
	⚙️	https://github.com/AmericasNLP/americanlp2021/pull/5
English en	☆	OPUS: Europarl, GlobalVoices, News-Commentary, TED2020, Tatoeba, Open-Subtitles (Tiedemann, 2012)
	📖	OPUS: bible-uedin (Christodoulopoulos and Steedman, 2015)
Spanish	📖	<i>spa-x-bible-americas, spa-x-bible-hablahoi-latina, spa-x-bible-lapalabra, spa-x-bible-newworld, spa-x-bible-nuevadehoi, spa-x-bible-nuevaviviente, spa-x-bible-nuevointernacional, spa-x-bible-reinavaleracontemporanea</i>

Table 4: Data resources used for training. 🌟 refers to the official training data provided by the organizers. ☆ marks datasets from the *extra* categories already used in 2021, and ★ refers to new *extra* data. 📖 designates Bible identifiers from the JHUBC. Datasets marked with ↻ are created using backtranslation, datasets marked with 🔄 using pivot translation from English to Spanish. Conversion tables and scripts are listed under ⚙️.